MUSTAFA BOZKAYA

Senior AI & Platform Engineer

info.mustafabozkaya@gmail.com +90 (545) 517 07 75 Istanbul, Turkiye linkedin.com/in/mustafa-bozkaya

PROFESSIONAL PROFILE

Innovative Senior AI and Platform Engineer with over 6 years of professional experience in software engineering, specializing in the design and implementation of scalable distributed systems and high-performance backend services. Expert in building production-grade platforms using Python and Go, with a deep focus on integrating Large Language Models (LLMs) and developing autonomous Agentic AI workflows. Proven track record in orchestrating end-to-end MLOps lifecycles, optimizing containerized microservices on Kubernetes, and architecting robust gRPC and RESTful APIs for high-traffic environments. Dedicated to bridging the gap between advanced AI research and reliable enterprise infrastructure while mentoring cross-functional engineering teams in Agile environments.

EXPERIENCE

Lead Data & AI Engineer

August 2025 - Present

Yaska Group

- Spearheading the architectural design and implementation of highly scalable distributed systems across multiple group subsidiaries, ensuring 99.9% uptime for core backend services.
- Developing high-quality backend services using Python and Go, focusing on efficient data processing and low-latency model inference responses.
- Architecting and deploying enterprise-grade Generative AI solutions and LLM integrations that automated internal workflows, resulting in a 25% increase in operational efficiency.
- Designing multi-step autonomous agent systems with complex reasoning capabilities, utilizing tool-calling and sophisticated state management techniques.
- Leading cloud-native deployment strategies on Kubernetes, optimizing CI/CD pipelines to facilitate rapid and reliable delivery across multiple cross-functional teams.
- Coordinating cross-functional engineering initiatives, aligning technical roadmaps with business goals and providing technical mentorship to junior backend engineers.
- Establishing rigorous data security and privacy protocols for the secure processing of corporate data within AI models, ensuring compliance with international standards.

SKILLS

Python (Advanced) Go Java
Distributed Systems Architecture
Agentic AI & Workflow Engineering
LLM Integration (GPT, LLaMA, DeepSeek)
RESTful & gRPC API Design
Kubernetes & Docker
Infrastructure as Code (Terraform)
CI/CD Pipelines (GitHub Actions, Jenkins)
Observability (Prometheus, Grafana, ELK)
Vector Databases (Pinecone, Chroma, Qdrant)
MLOps (MLflow, Kubeflow, n8n)
RAG (Retrieval-Augmented Generation)
Data Structures & Algorithms

EDUCATION

İstanbul Üniversitesi
Bachelor of Engineering - BE, Computer Engineering
2011 - 2014

İstanbul Üniversitesi
Bachelor of Engineering - BE, Industrial Engineering
2014 - 2018

LANGUAGES

English (Professional Working Proficiency)
Turkish (Native)

Co-Founder | Lead AI MLOps Engineer

July 2024 - Present

Premium AI

- Designing and managing the end-to-end MLOps lifecycle for products built on AGI, LLM, and Vision-Language Models (VLM), bridging the gap between research and production environments.
- Engineered autonomous agents using LangChain and LlamaIndex capable of performing complex task execution, tool selection, and multi-step reasoning.
- Built high-accuracy RAG (Retrieval-Augmented Generation) pipelines using Pinecone and Chroma, providing real-time contextual information to models with sub-second latency.
- Implemented advanced model quantization techniques including QLoRA and GGUF, reducing hardware overhead by 40% while maintaining high inference performance.
- Leveraged Terraform for Infrastructure as Code (IaC) to ensure reproducible and scalable AI infrastructure across hybrid cloud environments (AWS/GCP).
- Developed a comprehensive observability architecture using ELK Stack, Prometheus, and Grafana to monitor agentic decision-making processes and system health.
- Architected gRPC-based microservices to facilitate high-throughput communication between distributed AI components and backend modules.

AI || MLOps Engineer

July 2023 - July 2024

AI Planet

- Fine-tuned GPT-2 and LLaMA models via the Hugging Face ecosystem for specialized customer support bots, achieving a 15% improvement in response accuracy.
- Optimized model response times by 20% through the implementation of Qdrant-backed vector search and efficient retrieval strategies.
- Applied LoRA and 4-bit quantization to reduce production model footprints by 40% without compromising performance metrics.
- Developed containerized AI APIs using FastAPI capable of handling over 10,000 concurrent requests per second with high availability.
- Automated testing and deployment workflows using GitHub Actions and Jenkins, reducing time-to-market for new features by 40%.
- Managed large-scale Kubernetes clusters on Google Cloud Platform (GCP) to provide dynamic scaling for AI services under fluctuating traffic loads.

Artificial Intelligence R&D Engineer

May 2022 - June 2023

EPİK ROBOTİK

- Developed a computer vision-based navigation system for autonomous mobile robots using OpenCV, PyTorch, and ROS (Noetic/Humble), reaching 95% obstacle detection accuracy.
- Built NLP-based command recognition systems using BERT and Hugging Face Transformers to enable robots to interpret and execute complex verbal commands.
- Optimized inference speed by 30% through the application of CUDA acceleration and model distillation techniques for edge computing devices.
- Led R&D initiatives for sensor fusion and deep learning algorithm integration into robotic hardware platforms.
- Designed robust backend communication layers to handle telemetry data between robots and central control systems using Python-based microservices.

Software Engineer

February 2014 - November 2020

Freelancer.com

- Engineered scalable web applications and SaaS platforms using React, Next.js, and Node.js, focusing on modular architecture and clean code principles.
- Designed and optimized complex relational and non-relational database schemas using PostgreSQL and MongoDB for high-traffic e-commerce solutions.
- Implemented cloud infrastructure on AWS, managing containerized deployments and automated CI/CD pipelines for diverse international clients.
- Developed RESTful APIs and integrated third-party ERP systems to streamline business operations for enterprise-level clients.
- Authored comprehensive technical documentation and unit tests, ensuring high software quality and maintainability across long-term projects.